information retrieval tool, having found the birth date, would select the critique and produce a summary. This summary however will not actually contain the birth date of D. H. Lawrence as the author's birth date would be of almost no importance to the main topic in a critique of "Sons and Lovers". Nor would the summary identify where in the critique the information about the author's birth date appears.

Ins B3

According to a first aspect of the present invention there is provided apparatus for summarising data sets, the apparatus having:

an input for receiving a data set to be summarised;

sectioning means for dividing said received data set into one or more sections according to pre-determined criteria;

ranking means operable for each said section to compare data within the said section with one or more target data items and for calculating a ranking value for the said section, said ranking value being dependent on the outcome of said comparisons for the said section; and

selecting means for compiling a customised summary of the data set by selecting one or more of said one or more sections according to their respective ranking values.

For instance, sections having a ranking value which is above (or below, depending on the circumstances) a preselected threshold might be selected.

According to a second aspect of the present invention there is provided a method for generating a customised summary of a data set, the method including the steps of:

i) receiving, as input, a data set to be summarised;

ii) dividing said data set into sections according to predetermined criteria;

iii) comparing data items in each said section against one or more target data items;

iv) calculating a ranking value for each said section in dependence upon the outcome of the respective said comparisons; and

v) compiling a customised summary of said data set by selecting one or more of said one or more sections according to their respective ranking values.

Preferably, target data items can be loaded to the target data item store by a user, for instance either directly or via a user profile. An advantage of such embodiments of the invention is that they enable a summarising tool to generate a

Article 34 Amdt.



summary of a data set that includes target data items specified by a user for whom the summary is generated.

There are many additional features which may be provided, separately or in combination, by preferred embodiments of the present invention and at least some of these are discussed as follows.

Data sets may be divided into sections according to sentences, paragraphs, and other punctuation. Alternatively, other formats such as pages and chapters and headings may form section boundaries.

Within the context of summarising data sets, a key data item is a data item that forms a substantive component of the information contained within the data set. For example, in a document consisting of written prose, articles and conjunctions (for instance words such as 'it', 'are', 'as', 'the', 'when', 'they', 'by' etc.) are typically not considered to be key data items. This is because they do not identify subject matter contained within the data set.

According to preferred features of the present invention, the apparatus includes:

means for identifying one or more key data items in each said section according to a pre-determined stop list;

calculating means operable for each said section to calculate one or more distribution values, each said distribution value representing a different pre-determined measure of the distribution, in said data set, of key data items identified in the said section; and

adjustment means for adjusting said ranking value for each said section according to the respective said one or more distribution values.

Preferably the method includes the steps of:

a) identifying key data items within each said section from step ii) according to a pre-determined stop list;

b) calculating, for each said section, one or more distribution values each representing a pre-determined measure of the distribution of the key data items of the said section in said data set; and

c) adjusting said ranking value from step iv) for each said section in dependence upon the respective said one or more distribution values.

Article 34 Amdt



Refining ranking values according to the distribution of key data items within the data set allows the summary to detail target data items within the context of the main topic of the data being summarised. This increases the user's ability to determine how relevant a particular data set is for their intended purpose.





AMENDED SHEET